

Probability Lab 5: Q-Q plots

One way to test data to determine if the normal distribution is appropriate is to do what is called a Q-Q plot (this stands for quantile-quantile). Here is the idea. Suppose one makes a sequence of measurements of a random variable X . For definiteness let's suppose that the results are,

53, 55, 58, 53, 43, 41, 51, 42, 53, 54

The first step in making a Q-Q plot is to put these observations in increasing order. Here is how you can use R to do this. First define,

$x = c(53, 55, 58, 53, 43, 41, 51, 42, 53, 54)$

This creates a vector x with the appropriate entries (the c stands for “concatenate”—that is string the numbers together to make a vector). Now issue the R command,

$y = sort(x)$

If you look at y you find,

$y = (41, 42, 43, 51, 53, 53, 53, 54, 55, 58)$

The vector y has the same entries as x but arranged in increasing order. If we suppose that y is a normal random variable then for constants μ and σ we would have $y = \mu + \sigma Z$ where Z is the standard normal random variable. To get an idea if this is a plausible relation we estimate the value of Z which corresponds to a value for y in the following way. Consider the value 43 for y . The 10 observations of y divide up the possible values for y into 11 bins. There are 3 bins less than or equal to 43 so we estimate the probability that $y \leq 43$ as $3/11$. For our estimate of the normal random variable Z that corresponds to 43 for y we choose the number z so that $P(Z \leq z) = 3/11$. It is possible to use the normal table to estimate z but it is much simpler to use R. The syntax is,

$z = qnorm(3/11)$

The number z is called the $3/11^{th}$ quintile for the standard normal distribution. In this particular case $z = -.6046$. Our task is now to plot the values (z, y) for the observed values of y . If the graph “looks linear” this is a good indication that data is well fitted by a normal distribution. Note that the y intercept should be the mean and the slope should be the standard deviation. Here is how to use R to automate the whole process. You can generate all the quantiles at once by,

$z = qnorm(1 : 10/11)$

To see the graph issue the command

$plot(z, y)$

If you do it in this particular case you will find that you don't get a very good fit to a linear graph at all.

For homework I've given you three problems to find Q-Q plots for at the end of the chapter on normal random variables. In some cases you are not given the full string of observations for the random variable but instead you are given a value for the random variable say, 70, and the fraction, .56, of observations that are less than or equal to 70. In this case the point you want to plot for the Q-Q plot is $(qnorm(.56), 70)$.

For this lab I want you to simulate some random variables and use the Q-Q test to see if they are well approximated by a normal distribution. First let's look at the binomial distribution. Define,

$x = rbinom(1000, size = 50, prob = .5)$

This is a simulation of 1000 trials for a binomial random variable which counts the number of heads in 50 coin flips. To do the Q-Q plot for this data write,

$y = sort(x)$

and

```
 $z = qnorm(1 : 1000/1001)$ 
```

and

```
 $plot(z, y)$ 
```

Use the plot to estimate μ and σ . Repeat this experiment for $size = 100, 500$ and 1000 . Note that no random variable which has lots of repeated values can produce a linear graph (the repeats produce flat spots). In spite of this you do see something that looks linear, not so? Formulate a normal approximation to the binomial on the basis of this experiment—under what circumstances do you guess this approximation will be good.

Do the same experiment for the Poisson distribution for different choices for the mean $\lambda = 1, 10,$ and 50 . The syntax for generating 1000 trials for a Poisson random variable with $\lambda = 10$ is,

```
 $x = rpois(1000, 10)$ 
```

From this experiment formulate a normal approximation to the Poisson distribution. Under what circumstances do you think this approximation will be a good one?